Infraestructura de Acceso a Datos Primarios con Aporte de Semántica en Repositorios Digitales

Renato Mazzanti[1,2,3], Marcos Zarate[1,2,4], Gustavo Samec[1,2,3], Carlos Buckle[1,2]

[1] Departamento de Informática, Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB), Puerto Madryn, Chubut, Argentina.
[2] LINVI, Laboratorio de Investigación en Informática, Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB)
[3] Unidad de Gestión de la Información, Centro Nacional Patagónico, Consejo Nacional de Investigaciones Científicas y Técnicas (CCT CONICET-CENPAT)
[4] Centro para el Estudio de Sistemas Marinos, Centro Nacional Patagónico, Consejo Nacional de Investigaciones Científicas y Técnicas (CESIMAR-CENPAT-CONICET)
renato@cenpat-conicet.gob.ar , zarate@cenpat-conicet.gob.ar , gsamec@cenpat-conicet.gob.ar, cbuckle@unpata.edu.ar

Resumen

En la actualidad hay grandes retos científicos que necesitan analizar una gran cantidad y diversidad de datos, por ello, el acceso a los datos primarios producidos por las investigaciones está cobrando cada vez mayor relevancia. Compartir estos datos, brindar acceso a los mismos y permitir su reutilización genera innumerables ventajas para la comunidad científica. La Argentina cuenta con la ley 26.899 que promueve la creación de políticas y mecanismos para la gestión de datos primarios científicos a nivel nacional. Este trabajo, presenta un proyecto de investigación aplicada y desarrollo experimental que aportará modelos y componentes para las infraestructuras de Repositorios Digitales Institucionales (RDI) apuntando a facilitar el acceso a datos primarios. Hará uso de teorías y tecnologías de la web semántica para la definición de consultas integradas entre repositorios, como así también para la generación de productos de síntesis que faciliten la tarea de análisis y descubrimiento de conocimiento a los investigadores. El proyecto trabajará sobre las bases de datos de los Sistemas Nacionales de Datos Biológicos y del Mar y además pretende fortalecer el RDI de la Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB) actualmente en desarrollo.

Palabras clave: Datos Primarios, Repositorios Digitales, Web Semántica, Gestión de datos científicos.

1. Contexto

El proyecto Infraestructura de acceso a Datos Primarios con aporte de semántica en Repositorios Digitales se elaboró en el Laboratorio de Investigación en Informática (LINVI) de la UNPSJB y está integrado por docentes investigadores de la Facultad de Ingeniería Sede Puerto Madryn, con participación de estudiantes y graduados de las carreras de dicha Sede. Colaboran en el equipo de trabajo, docentes investigadores de la Universidad Nacional del Sur (UNS) y de la Universidad Nacional de San Luis (UNSL). El proyecto fue avalado por el Consejo Directivo de la Facultad de Ingeniería de la UNPSJB y será financiado por la Secretaría de Ciencia y Técnica de la Universidad para llevar a cabo durante 2018-2020. Para el desarrollo del proyecto se utilizará como repositorio documental el RDI actualmente en desarrollo en la UNPSJB. Dado que integrantes del proyecto desarrollan tareas científicas y de soporte en el Centro Científico Tecnológico CONICET-CENPAT, se tomará como referencia el RDI de dicho Centro [1] y el RDI CONICET Digital [2]. Respecto a repositorios públicos de datos primarios se utilizaran el Sistema Nacional de Datos Biológicos (SNDB) [3] y el Sistema Nacional de Datos del Mar (SNDM) [4].

2. Introducción

Recientemente se sancionó en Argentina la Ley 26.899 "Repositorios Digitales Institucionales de Acceso Abierto" que establece la obligatoriedad de desarrollar repositorios digitales, propios o compartidos, por parte de los organismos e instituciones públicas que componen el Sistema Nacional de Ciencia, Tecnología e Innovación. Esta ley requiere el establecimiento de políticas institucionales para la gestión, el acceso público y la preservación de datos primarios de investigación. En éste contexto la UNPSJB adhirió a ésta iniciativa y promovió la creación del RDI según la Ordenanza Nro. 159 del Consejo Superior.

Como ha sucedido en el resto del mundo, la mayoría de los RDIs implementados, alberga material documental, tradicionalmente asociados a la producción científica (artículos, disertaciones, reportes, libros, tesis, etc.) y un porcentaje muy bajo de datos científicos [5], lo que dificulta la posibilidad de vincular las publicaciones con los datos que le dan sustento. Las necesidades actuales de interoperabilidad, usabilidad y uniformidad de vocabularios han llevado a repensar los modelos clásicos de organización del conocimiento en RDIs. En este contexto es donde comienzan a adquirir relevancia las búsquedas semánticas, como herramienta para la interrelación inteligente entre repositorios abiertos. Incorporar contenido semántico permite que las búsquedas no sean sobre "bolsas de palabras" sino que sean búsquedas conceptuales que permitan el descubrimiento inteligente de conocimiento [6]. Por otra parte, es necesaria la producción de soluciones que acompañen y fortalezcan las normativas antes mencionadas sobre datos abiertos, en la cual gran parte de los datos científicos de investigaciones financiados con fondos públicos deben ser de dominio público y permanecer disponibles para otros proyectos que lo requieran. La posibilidad de incorporar metadatos a los conjuntos almacenados y definir modelos conceptuales permite la utilización de dichos datos a las generaciones futuras, independientemente de la supervivencia de los integrantes del proyecto que originaron los datos [7]. Este proyecto plantea la definición de modelos y componentes para las infraestructuras de RDI tendientes a facilitar el acceso y la explotación de datos primarios residentes en diferentes bases de datos. Para ello, hará uso de teorías y tecnologías de la web semántica, abordará la vinculación entre las publicaciones y los datos (data-citation) y aspectos relacionados con la generación de productos de síntesis que faciliten la tarea de análisis y descubrimiento de conocimiento a los investigadores. Como plataforma para el desarrollo de prototipos se utilizara el RDI de la UNPSJB, actualmente en desarrollo y dado que los integrantes del proyecto se especializan en la gestión de datos primarios biológicos y del mar, se tomarán como referencia los RDIs de CONICET Digital y del CCT CONICET-CENPAT y los repositorios de datos científicos de acceso público existentes de los Sistemas Nacionales: SNDB [3] y SNDM [4], ambos son iniciativa del Ministerio de Ciencia, Tecnología e Innovación Productiva (MINCYT) conjuntamente con el Consejo Interinstitucional de Ciencia y Tecnología (CICyT) enmarcada dentro del Programa de Grandes Instrumentos y Bases de Datos.

3. Motivación

La mayor parte de las universidades y centros de investigación de todo el mundo, incluido nuestro país, disponen ya de repositorios institucionales que almacenan los resultados de investigación de sus miembros (principalmente artículos, comunicaciones a congresos y tesis doctorales), en este sentido la Ley Nacional 26.899 establece que los datos primarios de las investigaciones científicas financiadas con fondo públicos deben estar disponibles para que los usuarios de éste tipo de material puedan en forma gratuita, leer, descargar, copiar, distribuir, imprimir, buscar o enlazar los textos completos de los artículos científicos, y usarlos con propósitos legítimos ligados a la investigación científica en consonancia con el modelo de Acceso Abierto (AA) [7].

En la actualidad existen limitaciones para lograr los objetivos que promueve el AA. Un análisis crítico de la literatura disponible indica que hay limitaciones relacionadas con:

1. La calidad de los datos, ya que hay errores que podrían poner en entredicho los resultados de una investigación [8, 9].

2. Integración de los datos primarios de diferentes investigaciones dado que la forma de relacionarlos carece de vocabularios controlados o términos comunes que puedan ser relacionados.

3. Existe un alto porcentaje de la información que describe éstos datos primarios (metadatos) que suele ser semiestructurada [10]. Desde el punto de vista semántico, esto significa que puede existir un conocimiento implícito sobre éste contenido el cual no es explotado para descubrir nuevas relaciones entre ellos.

Al evaluar un conjunto de datos que es de una fuente desconocida, un usuario se basa fundamentalmente en las propiedades visibles del conjunto de datos, tales como título, editor, tamaño del conjunto de datos, etc. Los aspectos relativos a la calidad de los datos generalmente quedan fuera de las posibilidades de evaluación quedando en el terreno de la incertidumbre. Se pueden mencionar como antecedentes que en GBIF España[1] en 2007 se hicieron intentos de corregir ésta falencia mediante la implementación de un índice que permita a un usuario evaluar la calidad del conjunto de datos Darwin Core[2] que ellos publican y de éste modo hacer un seguimiento de las mejoras en la calidad de datos a lo largo del tiempo. Se trataba de un índice explícito simple el que fue mejorado en 2010 denominándose *Índice de Calidad Aparente* (ICA) [11]. ICA resume la calidad de los datos en las tres dimensiones para los datos de biodiversidad: taxonómica, geoespacial y temporal. Para esta temática el proyecto tendrá en cuenta recomendaciones del Grupo de Interés de Calidad de Datos de Biodiversity Information Standards (TDWG 2016).

De acuerdo con [12] el surgimiento de la web semántica ha permitido obtener soluciones a problemas habituales en la búsqueda de información gracias a la utilización de una infraestructura común, mediante la cual, es posible compartir, procesar y transferir información de forma sencilla. Entre los principales trabajos podemos encontrar [13] en el cual los autores proponen una metodología y una implementación capaz de explotar metadatos de un repositorio implementado con DSpace[3], haciendo uso de los beneficios de la web semántica para descubrir y vincular información. En el caso de [9] los autores describen los avances logrados en el desarrollo del repositorio digital Academic Commons (AC), de la Universidad de Columbia, como un repositorio interoperable, a través del uso de RDF y tecnologías de Web Semántica.

4. Líneas de Investigación, Desarrollo e Innovación

Este proyecto tiene como línea principal de investigación el Modelado Conceptual en la Web Semántica [14] para definir componentes que faciliten la integración de datos primarios y la explotación de información útil para los investigadores [15]. Además de ésta área, se abordarán otras relacionadas con la definición de infraestructuras para e-ciencia [16], plataformas para desarrollo de RDIs interoperables distribuidos e inteligencia para el descubrimiento de información.

5. Resultados esperados

El objetivo general del proyecto es definir componentes de arquitectura para ser implementados en RDIs como herramientas para facilitar el acceso unificado entre las publicaciones y los datos primarios que las

[1] http://www.gbif.es/

[2] http://rs.tdwg.org/dwc/

[3] http://www.dspace.org/

sustentan y potenciar el descubrimiento de información útil para los investigadores.

Para ello, se relevará el estado actual respecto a la integración entre RDIs públicos de referencia y bases de datos científicos abiertos (particularmente datos biológicos) y se estudiarán teorías, tecnologías y recomendaciones definidas para la web semántica, para definir finalmente, modelos conceptuales de integración entre repositorios. También se abordarán estudios relacionados a las plataformas para publicación de datos primarios, a la implementación de RDIs, a los mecanismos de referencias (citas) y al descubrimiento inteligente de información científica.

Como desarrollo experimental, se utilizará el RDI de la UNPSJB, actualmente en desarrollo, sobre el cual se esperan alcanzar prototipos con características de usabilidad. Inicialmente, se realizó un relevamiento de las tecnologías disponibles para nuestra arquitectura, y se puntualizó sobre las plataformas D2RQ [17], Jena [18], OpenRefine [19] y GraphDB [20], entre otras. De este análisis, se determinó que las plataformas más convenientes para un primer prototipo son OpenRefine y GraphDB. El primero soporta extensiones para la creación de triplas RDF a partir de una gran variedad de formatos de entrada, tales como CSV, hojas de cálculo, JSON y el mismo formato RDF. Además, permite explorar y depurar los conjuntos de datos, aplicar transformaciones y definir vocabularios asociados a los diferentes campos, de una manera amigable. Por otro lado, GraphDB es un repositorio semántico que permite almacenar las tripletas generadas por OpenRefine y también trabajar con motores de inferencia y de consultas SPARQL [21] sobre estos datos estructurados.

6. Formación de recursos humanos

En éste proyecto participan seis docentes del Departamento de Informática de la Facultad de Ingeniería de la UNPSJB Sede Puerto Madryn. Dos de ellos están realizando carreras de doctorado y otras dos carreras de especialización y maestrías. Uno de los autores de éste trabajo está inscripto en el Doctorado en Ciencias de la Computación en la Universidad Nacional del Sur y cuenta con beca interna doctoral del CONICET hasta el 2020. También forman parte del proyecto 1 (un) graduado de la carrera de Licenciatura en Informática y 6 (seis) alumnos del ciclo superior. Su participación tiene el objeto de introducirlos a la tarea científica y permitirles incorporar conocimientos sobre temas no desarrollados en la currícula de la carrera. En el caso de los alumnos que están próximos a graduarse, la intención del proyecto es guiarlos en el desarrollo de sus tesinas en ésta rama de la disciplina. Otro aporte para la formación académica radica en la posibilidad que los alumnos puedan realizar Instancias Supervisadas de Formación en la Práctica Profesional en el marco de este proyecto.

Referencias

[1] Repositorio Institucional CCT CONICET-CENPAT, Ministerio de Ciencia Tecnología e Innovación Productiva. Presidencia de la Nación Argentina. http://www.repositorio.cenpat-conicet.gob.ar Accedido: 2018-02-15.

[2] Repositorio Institucional CONICET Digital, Ministerio de Ciencia Tecnología e Innovación Productiva. Presidencia de la Nación Argentina http://ri.conicet.gov.ar Accedido: 2018-02-15.

[3] Sistema Nacional de Datos Biológicos, Ministerio de Ciencia Tecnología e Innovación Productiva. Presidencia de la Nación Argentina. http://www.sndb.mincyt.gob.ar Accedido: 2018-02-15

[4] Sistema Nacional de Datos del Mar, Ministerio de Ciencia Tecnología e Innovación Productiva. Presidencia de la Nación Argentina. http://www.datosdelmar.mincyt.gob.ar Accedido: 2018-02-15.

[5] Silvia Arano, Gemma Martínez, Marina Losada, Marta Villegas, Anna Casaldaliga y Nuria Bel. La comunidad, Recursos y datos primarios de la universitat pompeu fabra: los repositorios institucionales como infraestructuras científicas: estudio de caso. Revista española de Documentación Científica, 34(3):385– 407, 2011.

[6] D. C. Li, H. Liu, C. G. Chute, and S. R. Jonnalagadda. Towards assigning references using semantic, journal and citation relevance. In 2013 IEEE International Conference on Bioinformatics and Biomedicine, pages 499–503, Dec 2013.

[7] Sarah CALLAGHAN. Preserving the integrity of the scientific record: data citation and linking. Learned Publishing, 27(5):S15–S24, 2014.

[8] Ben R. Martin. Whither research integrity? Plagiarism, self-plagiarism and coercive citation in an age of research assessment. Research Policy, 42(5):1005 – 1014, 2013.

[9] D. Fanelli. How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. Plos One, 4(5):e5738, 2009-05-29 00:00:00.0.

[10] Jung-Ran Park. Metadata quality in digital repositories: A survey of the current state of the art. Cataloging & Classification Quarterly, 47(3-4):213–228, 2009.

[11] Francisco Pando. Quantifying quality: the apparent quality index", a measure of data quality for occurrence datasets. Biodiversity Information Science and Standards, 1:e20533, 2017.

[12] Rob Koper. Use of the semantic web to solve some basic problems in education. 6, 07 2004.

[13] Dimitrios A. Koutsomitropoulos, Geor-gia D. Solomou, and Theodore S. Papatheodorou. Semantic query answering in digital repositories: Semantic search v2 for dspace. Int. J. Metadata Semant. Ontologies, 8(1):46–55, May 2013.

[14] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. Scientific American, 284(5):34–43, May 2001.

[15] Tony Hey, Stewart Tansley, and Kris-tin Tolle. The Fourth Paradigm: Data-Intensive Scientific Discovery. October 2009.

[16] Giuseppe Andronico, Valeria Ardizzone, Roberto Barbera, Bruce Becker, Riccar-do Bruno, Antonio Calanducci, Diego Carvalho, Leandro Ciuffo, Marco Farget-ta, Emidio Giorgio, Giuseppe Rocca, Al-berto Masoni, Marco Paganoni, Federi-co Ruggieri, and Diego Scardaci. e-infrastructures for e-science: A global view. 9:155–184, 06 2011.

[17] Christian Bizer and Andy Seaborne. Treating non-rdf databases as virtual rdf graphs.

[18] Brian McBride. Jena: A semantic web toolkit. IEEE Internet computing, 6(6):55–59, 2002.

[19] Ruben Verborgh and Max De Wilde. Using OpenRefine. Packt Publishing Ltd, 2013.

[20] Ralf Hartmut Guting. Graphdb: Mode-ling and querying graphs in databases. In VLDB, volume 94, pages 12–15, 1994.

[21] Eric Prud, Andy Seaborne, et al. Sparql query language for rdf. 2006.